Network of Epistatic Interactions Within a Yeast snoRNA

Title: Network of Epistatic Interactions Within a Yeast snoRNA

Authors: Olga Puchta[1], Botond Cseke[2]†, Hubert Czaja[3], David Tollervey[4,5], Guido Sanguinetti[2,4], Grzegorz Kudla[1]*

Affiliations:

[1]MRC Human Genetics Unit, IGMM, University of Edinburgh, Scotland, UK.

[2]School of Informatics, University of Edinburgh, Scotland, UK.

[3]Scott Tiger SA, Warsaw, Poland.

[4]SynthSys, Centre for Synthetic and Systems Biology, University of Edinburgh, Scotland, UK

[5]Wellcome Trust Centre for Cell Biology, University of Edinburgh, Scotland, UK.

*Correspondence to: Grzegorz Kudla
MRC Human Genetics Unit
University of Edinburgh
Crewe Road, Edinburgh EH4 2XU
Scotland, United Kingdom
work: +44 (0) 0131 651 8628
e-mail: gkudla@gmail.com

†current address: Microsoft Research Cambridge, Cambridge, UK

Abstract (122 words): Epistatic interactions play a fundamental role in molecular evolution, but little is known about the spatial distribution of these interactions within genes. To systematically survey a model landscape of intragenic epistasis, we quantified the fitness of ~60,000 *Saccharomyces cerevisiae* strains expressing randomly mutated variants of the 333-nt long U3 snoRNA. The fitness effects of individual mutations were correlated with evolutionary conservation and structural stability. Many mutations had small individual effects, but large effects in the context of additional mutations, indicating negative epistasis. Clusters of negative interactions were explained by local thermodynamic threshold effects, whereas positive interactions were enriched among large-effect sites and between base-paired nucleotides. We conclude that high-throughput mapping of intragenic epistasis can identify key structural and functional features of macromolecules.

One Sentence Summary: A high-throughput fitness assay identifies the complete network of epistatic interactions within a yeast RNA, and reveals mechanisms of epistasis.

Main Text: The effect of a mutation on phenotype may depend on the presence of additional mutations. This phenomenon, known as epistasis, explains synthetic lethal interactions, where a combination of two individually viable mutations causes death, and compensatory interactions, where the fitness cost of a mutation is reduced by a second mutation (*1, 2*). Epistasis plays a major role in evolution; it determines the accessibility of mutational pathways (*3*), thereby influencing the rate of adaptation and the diversity and robustness of genetic variants (*4, 5*). Although genome-wide studies have revealed a network of intergenic epistasis (*6*), it has been suggested that interactions within genes may be even more common (*7-11*). However, previous studies focused on relatively small networks of interactions, and the comprehensive pattern of epistasis has not yet been determined for any gene.

We used "doped" oligonucleotides to synthesize ~130,000 randomly mutated variants of the 333-nucleotide *Saccharomyces cerevisiae* gene SNR17A, which encodes the U3 small nucleolar RNA (snoRNA). U3 basepairs to the primary rRNA transcript (pre-rRNA) and this interaction is required for pre-rRNA cleavage and 18S rRNA biogenesis. Our mutagenesis approach ensures uniform coverage of mutations among positions 7-333, encompassing 98% of the gene, and prevents bias towards specific types of mutations (Figs. 1A, S1). We generated two independent mutant libraries, which contained on average 3 and 10 single nucleotide polymorphisms (SNPs) per allele, respectively. In addition to the SNPs, 43.6% of variants also contained short deletions (median length 1 nt) or insertions (median length 1 nt). All 981 (3*327) possible point mutations were represented in the library, and 99.4% of the 53,301 (327*326/2) possible pairs of sites were jointly mutated, most of them in alleles that contained additional mutations. To facilitate unambiguous identification of variants by high-throughput sequencing, we tagged each variant

with a unique 20-nt barcode (Fig. 1A) placed in a non-transcribed region downstream of the U3 gene to minimize interference with function.

To measure fitness, we used the D343 yeast strain, which contains a single copy of the wild-type U3 gene under control of a galactose-inducible promoter (*12*). D343 cells can grow in galactose-containing medium, but shifting to glucose results in downregulation of U3 and growth arrest. Transformation of wild-type U3 on a plasmid allows the cells to survive on glucose, but non-functional U3 mutants do not support growth (Fig. S2). We transformed D343 cells with centromeric plasmids carrying the U3 mutant libraries, and measured the frequency of each mutant during competitive growth on glucose (Fig. 1B). As expected, non-functional variants decreased in frequency during the competition, whereas the wild-type gene increased (Fig. 1C). Growth patterns were reproducible between four replicate experiments and across replicate U3 variants within an experiment (Figs. 1D, S3).

We measured the logarithm of relative fitness (log fitness) of ~60,000 variants that passed quality filters, by fitting exponential decay curves to the barcode count data (*13*). Log fitness of wild-type U3 was set to 0. We first focused on the effects of single mutations in an otherwise wild-type gene (*13*). In most positions, mutations were tolerated with minimal effect on fitness (Fig. 2A). The exceptions were the conserved protein binding sites known as Box B, C, C' and D, mutations in which are lethal or highly deleterious. In addition, a moderate fitness decrease was observed for mutations within stems I, II, III and VI, particularly in G-C base pairs located at the base of stems, suggesting a role in structural stability. Folding predictions confirmed that destabilizing mutations in individual stems reduced fitness (Fig. S4). The 5nt 3'-terminal stem of U3 confers protection from degradation by 3'-5' exonucleases (*12*). Mutations in this stem reduced fitness proportionally to their predicted effect on RNA folding strength (Figs. S4, S5).

U->C mutations in positions 178 and 191 were highly deleterious (Fig. S5), possibly because they created consensus binding motifs (UCUUG) for the RNA degradation factor Nab3 (*14*). The fitness effects of mutations were slightly larger at 37C compared to 30C, consistent with the destabilizing effect of temperature on U3 structure (Fig. S6). We found no mutations that consistently increased fitness, suggesting that wild-type U3 is optimally adapted for function. In conditions where the genomic copy of U3 was coexpressed with the mutant library, the mutations had no effect, indicating lack of dominant negative or gain of function effects (Fig. 1D). Overall, these results show the expected pattern of single-site fitness effects and support the reliability of the measurements.

We then calculated p_i, the aggregate log fitness effect of position i across all genetic backgrounds represented in the library (*13*). In contrast to the single mutants (Fig. 2A), most positions showed substantial effects on fitness in combination with other mutations (Fig. 2B). The variation in p_i across the gene was reproducible between replicate experiments, was observed in both mutant libraries, was robust to the exclusion of outlier variants, and did not reflect the co-occurrence of mutations with large-effect mutations in other sites (Figs. S6,7). Most positions with very negative p_i map to the conserved core of the U3 gene (stems I-III, 5' hinge, 3' hinge, Fig. 2B). In contrast, sites with near-zero p_i were located in the fungal- or yeast-specific regions of the molecule (stems IV-VI, Fig. 2B). The p_i values were correlated with fitness effects in wild-type background (Fig. 2C, Spearman rho=0.57, pval$<2*10^{-16}$). However, the relationship was markedly non-linear, indicating that many mutations had small effects in an otherwise wild-type U3 but large effects in the context of additional mutations.

This pattern suggests a high prevalence of negative epistatic interactions within U3. To confirm this, we analyzed the distribution of fitness effects as a function of the number of mutations in

each allele (Fig. 2D). As expected, the average fitness of variants decreased with increasing numbers of mutations. Notably, measured fitness was consistently lower than expected under an additive model (Fig. 2E). This indicates overall enrichment of negative epistatic interactions relative to positive interactions.

We estimated the strength of all pairwise interactions from measurements of single, double and multiple mutants (*13*) with a regression model (*15-17*) that explained approximately 86% of variance in measured fitness and produced similar patterns of interactions when applied to our replicate experiments (Fig. S9). We obtained a consensus set of epistatic interactions by averaging the interaction estimates from all four glucose competition experiments. Plotting these interactions resulted in a characteristic tartan pattern (Fig. 3A), in which several positions in the gene showed strong positive interactions with most other sites. These hubs of positive epistasis correspond to the C', C and D boxes, which are highly conserved in evolution and show the largest individual effects on fitness. This observation indicates a saturation effect, whereby large-effect mutations inactivate the gene to such an extent that additional mutations become irrelevant, resulting in positive epistasis. Consistently, sites with large individual effects showed a strong bias towards positive epistasis (Wilcoxon test, $p<2*10^{-16}$; Fig. 3B,C), and the fitness of C', C and D box mutants did not depend strongly on the presence of additional mutations in the gene. The saturation effect is the within-gene equivalent of positive epistasis between pairs of mutations that independently inactivate the same metabolic pathway (*18, 19*).

We also expected positive interactions between base-paired positions, due to the presence of compensatory mutations, which are common in RNA evolution (*5*). Indeed, base-paired positions showed an enrichment of positive epistasis relative to all pairs of positions (Wilcoxon test, $p=2*10^{-7}$; Fig. 3D). In particular, all positions within the essential terminal stem formed strong

positive interactions with their corresponding base-paired residues (Fig. 3E). To test whether positive epistasis can be used to predict RNA folding, we intersected the set of positive interactions with a list of all potentially interacting triplets of nucleotides (*13*). In this analysis, 5 of 6 of the strongest positive interactions corresponded to known basepairs. When we used these interactions as constraints in RNA folding prediction, the accuracy of predicted secondary structure was improved (Fig. S10).

Despite the enrichment of positive interactions among large-effect or base-paired sites, negative interactions were more common overall (Fig. 3A,B). Whereas the strongest positive interactions typically involved at least one large-effect position (Fig. 4A), negative interactions were common among low- and intermediate-effect sites, and were distributed throughout the molecule, with enrichment in the conserved core (Fig. 4B). The strength of negative epistasis was inversely correlated with the distance along the primary sequence: 8 out of 10 strongest negative interactions were between pairs of adjacent nucleotides, and the median distance between the 100 most strongly interacting pairs was 18 nt. We thus focused on a hotspot of interactions encompassing the 3' hinge (Fig. S11A) that mediates base-pairing between U3 and the pre-ribosomal RNA (pre-rRNA), and is necessary for the pre-rRNA cleavage step of ribosome biogenesis (*20*). Our results suggest that the 3’ hinge can tolerate a single SNP, but that multiple mutations within this region reduce fitness, probably because they disrupt U3-rRNA binding. A similar, but less pronounced pattern was found in the 5’ hinge area. Our analysis shows that the thermodynamic threshold model, wherein fitness decreases abruptly when molecule stability falls below a certain level (*9*), also operates at the level of interactions between distinct molecules (Figure S11B).

In genome-wide studies, epistatic interactions between genes correlate with physical contacts, coevolution, and co-occurrence within biochemical pathways (*6*). Mapping genetic interaction networks therefore provides information about cellular organization. We postulate that intragenic interaction maps will similarly illuminate patterns of molecular organization. This and other studies (*8, 17, 21, 22*) suggest that within-gene epistatic interactions are enriched among residues in physical proximity. Were this correlation sufficiently strong, intra-gene epistasis would identify secondary and tertiary structures of macromolecules. Notably, recent studies have successfully predicted the 3D structures of proteins and complexes by measuring coevolution between residues within protein alignments, a phenomenon intimately linked to epistasis (*23, 24*). Improved methods to extract structurally relevant interactions from the dense network of intramolecular epistasis should allow macromolecular structures to be derived from maps of within-gene epistasis.

References and Notes:

1. P. C. Phillips, Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 9, 855 (Nov, 2008).
2. J. A. de Visser, J. Krug, Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet* 15, 480 (Jul, 2014).
3. D. M. Weinreich, N. F. Delaney, M. A. Depristo, D. L. Hartl, Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312, 111 (Apr 7, 2006).
4. J. I. Jimenez, R. Xulvi-Brunet, G. W. Campbell, R. Turk-MacLeod, I. A. Chen, Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proc Natl Acad Sci U S A* 110, 14984 (Sep 10, 2013).
5. M. V. Meer, A. S. Kondrashov, Y. Artzy-Randrup, F. A. Kondrashov, Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature* 464, 279 (Mar 11, 2010).
6. M. Costanzo *et al.*, The genetic landscape of a cell. *Science* 327, 425 (Jan 22, 2010).
7. B. Lehner, Molecular mechanisms of epistasis within and between genes. *Trends Genet* 27, 323 (Aug, 2011).
8. B. H. Davis, A. F. Poon, M. C. Whitlock, Compensatory mutations are repeatable and clustered within proteins. *Proceedings. Biological sciences / The Royal Society* 276, 1823 (May 22, 2009).
9. S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, D. S. Tawfik, Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* 444, 929 (Dec 14, 2006).
10. C. Bank, R. T. Hietpas, J. D. Jensen, D. N. Bolon, A systematic survey of an intragenic epistatic landscape. *Mol Biol Evol* 32, 229 (Jan, 2015).
11. A. I. Podgornaia, M. T. Laub, Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* 347, 673 (Feb 6, 2015).
12. D. A. Samarsky, M. J. Fournier, Functional mapping of the U3 small nucleolar RNA from the yeast Saccharomyces cerevisiae. *Mol Cell Biol* 18, 3431 (Jun, 1998).
13. Materials and methods are available as supporting material on Science Online.
14. W. Wlotzka, G. Kudla, S. Granneman, D. Tollervey, The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. *EMBO J* 30, 1790 (May 4, 2011).
15. J. Otwinowski, I. Nemenman, Genotype to phenotype mapping and the fitness landscape of the E. coli lac promoter. *PLoS One* 8, e61570 (2013).
16. J. Otwinowski, J. B. Plotkin, Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proc Natl Acad Sci U S A* 111, E2301 (Jun 3, 2014).
17. T. Hinkley *et al.*, A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat Genet* 43, 487 (May, 2011).
18. R. P. St Onge *et al.*, Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* 39, 199 (Feb, 2007).
19. X. He, W. Qian, Z. Wang, Y. Li, J. Zhang, Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nat Genet* 42, 272 (Mar, 2010).
20. L. M. Dutca, J. E. Gallagher, S. J. Baserga, The initial U3 snoRNA:pre-rRNA base pairing interaction required for pre-18S rRNA folding revealed by in vivo chemical probing. *Nucleic Acids Res* 39, 5164 (Jul, 2011).
21. H. Braberg *et al.*, From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. *Cell* 154, 775 (Aug 15, 2013).
22. C. Li, The fitness landscape of a tRNA gene. *submitted*, (2015).
23. T. A. Hopf *et al.*, Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3, (2014).
24. D. S. Marks *et al.*, Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 6, e28766 (2011).
25. M. Beltrame, D. Tollervey, Base pairing between U3 and the pre-ribosomal RNA is required for 18S rRNA synthesis. *EMBO J* 14, 4350 (Sep 1, 1995).
26. A. Stotz, P. Linder, The ADE2 gene from Saccharomyces cerevisiae: sequence and new vectors. *Gene* 95, 91 (Oct 30, 1990).
27. R. D. Gietz, R. A. Woods, Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods in enzymology* 350, 87 (2002).
28. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357 (Apr, 2012).

29. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078 (Aug 15, 2009).
30. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J Mol Biol* 215, 403 (Oct 5, 1990).
31. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841 (Mar 15, 2010).
32. H. A. Orr, Fitness and its role in evolutionary genetics. *Nat Rev Genet* 10, 531 (Aug, 2009).
33. B. Charlesworth, D. Charlesworth, *Elements of Evolutionary Genetics*. (Roberts and Company Publishers, 2010).
34. P. Mccullagh, Generalized Linear-Models. *Eur J Oper Res* 16, 285 (1984).
35. T. Minka, Expectation propagation for approximate Bayesian inference. *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence*, (2001).
36. R. Tibshirani, Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met* 58, 267 (1996).
37. A. J. Saldanha, Java Treeview--extensible visualization of microarray data. *Bioinformatics* 20, 3246 (Nov 22, 2004).
38. K. Darty, A. Denise, Y. Ponty, VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25, 1974 (Aug 1, 2009).
39. M. Krzywinski *et al.*, Circos: an information aesthetic for comparative genomics. *Genome Res* 19, 1639 (Sep, 2009).
40. R. P. Smyth *et al.*, Mutational interference mapping experiment (MIME) for studying RNA structure and function. *Nat Methods* 12, 866 (Sep, 2015).
41. I. L. Hofacker, Vienna RNA secondary structure server. *Nucleic Acids Res* 31, 3429 (Jul 1, 2003).

Acknowledgments: We thank B. Charlesworth, A. DeLuna, J. Plotkin, C. Schneider and J. Zhang for discussion; E. Wacker and A. Gallacher for technical assistance; M. Bartosovic for computational help; Edinburgh Genomics, Edinburgh Clinical Research Facility and IGMM technical support for next-generation sequencing. The sequencing and barcode count data was deposited in GEO (accession number GSE77709), and the data modeling pipeline can be found on https://github.com/terembura/EpistaticInteractionsYeastU3. GK and OP designed the project, OP performed experiments, OP, BC, HC, DT, GS and GK analysed the data, BC and GS designed and BC implemented the data modeling pipeline, GK wrote the manuscript with input from other co-authors. OP and GK were supported by Wellcome Trust grant 097383 and by the MRC, DT was supported by Wellcome Trust grant 077248, GS was supported by the ERC grant MLCS306999

Figure 1. Experimental mapping of the U3 fitness landscape.

A. To perform saturation mutagenesis of U3 we used PCR to assemble overlapping "97:1:1:1-doped" and non-doped oligonucleotides covering the whole length of the gene (*13*) and attached a unique 20-nt barcode to each variant. **B**. We cloned the U3 mutant library into centromeric plasmids and transformed the plasmids into the D343 yeast strain. **C**. Normalized read-counts from barcode sequencing for 7 randomly chosen wild type U3 variants (red) and 7 variants carrying a single mutation in box D (blue), in control (galactose) and competitive (glucose) conditions. Fitness was approximated by fitting exponential decay curves to the barcode count data. **D**. Rows indicate positions along U3 and columns indicate substitution to one of 4 bases: A, C, G and T. Log fitness effects are shown in blue for deleterious effects, red for positive effects and white for no effect on galactose in 30C (Gal), and glucose (Glu) in 30C and 37C. Genetic variants with one or two mutations were included in the analysis (*13*). Positions for which no data were obtained are shown in grey.

Figure 2. Distribution of fitness effects mapped to secondary structure.

A. The log fitness of single mutants at each position of U3 (f_i) is represented according to the colour scale, from blue for no effect, to red for effects -1 and stronger; non-mutagenized positions are white. **B**. The population-average log fitness effect for each position in the background of multiple mutations (p_i) (*13*). Evolutionarily conserved motifs are indicated on the secondary structure. **C**. Non-linear relationship between the fitness effects of single mutations in wild-type background (f_i) and in mutated backgrounds (p_i). **D**. Cumulative distributions of log fitness for mutants grouped by number of mutations per variant. **E**. Mean measured log fitness (white boxes) is always lower than expected in the absence of epistasis (grey boxes). Inclusion of epistasis improves the fit between the model and the data (dark grey boxes). The boxes show the median and inter-quartile ranges.

Figure 3. Localization of negative and positive epistasis.

A. Estimated pairwise epistatic interactions (w_{ij}) between positions in U3 (*13*), averaged between all experiments in glucose media. Negative interactions are green and positive are red; evolutionarily conserved motifs are indicated on the right and positions in U3 on the bottom. Data for the first 6 positions are skipped due to lower coverage of mutations (see Fig. S1). **B-D**. Distribution of w_{ij} showing negative epistasis (green bars) and positive epistasis (red bars) for all interactions (B), for sites with large individual effects (effect size <-1 for at least one site in pair, C), and among pairs of base-paired positions (D). **E**. Distributions of w_{ij} for individual positions in the terminal stem (75-78, 330-333). Red asterisks indicate interactions between known basepairs.

Figure 4. Network of epistatic interactions.

Circos plots showing patterns of 200 strongest positive (**A**) and negative (**B**) interactions within U3. Evolutionarily conserved motifs are indicated.

Supplementary Materials:

Materials and Methods

Figures S1-S11

Tables S1-S2

Materials and Methods:

U3 mutant library construction (saturation mutagenesis)

The library of U3 mutants was constructed in a two-step PCR approach that included assembly and amplification PCR reactions. To generate the "Big" library we performed assembly PCR with 6 overlapping, “doped” oligonucleotides (IDT, all sequences in Table S2). In the doped oligonucleotides, each position contained the wild-type nucleotide at 97% frequency and a 1%:1%:1% mix of the three other nucleotide types. This protocol resulted in a 3% mutation rate per position. To generate the "Small" library we performed 6 independent assembly reactions, each using 1 doped and 5 non-doped oligonucleotides, and we mixed the assembly PCR products in equimolar ratios. The Small library had an approximately 1% mutation rate per position. The assembly PCR products were used as a template for amplification PCR, to add 20-nucleotide random barcodes and restriction sites to each variant (primers: U3_start_SalI and U3_end_bar_EcoR, in Table S2). Nucleotides 1-6 of the U3 gene represent a SalI restriction site, and we did not introduce any mutations in this fragment. The product was cloned using SalI and EcoRI sites into the pU3-empty vector, under the control of native U3 promoter. Plasmid pU3-empty was constructed by replacing wild-type U3 coding sequence and 70 downstream nucleotides with 40 nt double stranded oligonucleotide Sal_oligo_Eco (Table S2) in the ARS-CEN vector pU3-wt, carrying an ADE2 marker (*25*), created on the backbone of pASZ11 (*26*).

Ligation products were transformed into DH5a/TOP10 competent cells and plated on 450 plates (~400 colonies per plate), for a total of 180,000 colonies. Colonies were pooled in PBS and used directly for MIDIprep (18 columns) (Qiagen, UK) to generate the U3 mutant plasmid library. MiSeq 300-nt paired-end sequencing of the entire insert (U3 mutated sequence, non-mutated linker and barcode) was performed to check the complexity of library and to associate random barcode sequences with U3 sequences.

Yeast strains

The D343 yeast strain (leu2, ura3, ade2, can1, his1, his3, trp1, U3aΔ, UASGAL:U3A::URA3, U3B::LEU2), where the wild type genomic version of U3 is expressed under the control of galactose promoter, was transformed with 80 μg of U3 mutant plasmid library using LiAc/SS-DNA/PEG High Efficiency Transformation Protocol (*27*). The whole transformation mix was transferred to 1l of liquid synthetic medium without adenine, supplemented with 2% galactose. To estimate the efficiency of transformation, 100ul of this liquid culture was plated just after inoculation and incubated for 2 days in 30C (based on this, the efficiency of the main transformation was usually 6250 colonies / 1 μg DNA). The liquid culture was grown in 30C until it reached OD600=2.0, diluted to OD600=0.2 (at least $5x10^8$ cells were transferred) and grown again to OD=1.0. This stage we called population G0 and performed HiSeq Illumina sequencing of barcodes (Edinburgh Genomics, UK) to check the complexity of the library.

Assigning U3 sequences to random barcodes

To identify the sequences of U3 variants associated with random barcodes, we used the plasmid library to amplify a ~400nt fragment containing the U3 mutated sequence, non-mutated linker

and random 20-nt barcode (primers: IndexX_PCR_U3_seq and R_PCR_U3bar_seq) and used this as a template for 300-nt paired-end MiSeq Illumina sequencing (Edinburgh Genomics, UK), using the Custom_Read1_seq_primer_U3 and Illumina Read2_sequencing_primer (Table S2). Candidate barcode sequences were extracted from between flanking sequences using blastall and bedtools programs. A large proportion of these candidate barcodes represented Illumina sequencing errors and were only present at very low frequency in the sequencing. To identify bona-fide barcodes associated with the U3 variants, we filtered the list of candidate barcodes according to their frequency in several sequencing runs (the original U3 mutant plasmid library and the G0 yeast population). To avoid errors in barcode assignment, we also filtered barcodes by similarity to other barcode sequences. We then used bowtie2 (*28*) and samtools (*29*) to map all reads to wild type U3 and to identify the consensus U3 sequence corresponding to each random barcode. We called a mutation if the same change at a given position appeared in at least 80% of reads with a given barcode. To avoid ambiguous assignments of barcodes to the U3 sequence, we discarded all barcodes for which we found any mutation in U3 present in between 20-79% of reads. We also discarded barcodes for which the linker sequence between U3 and the barcode was mutated, and barcodes for which the U3 sequence was not fully covered by reads.

Competition experiments

The competition experiments were performed in synthetic medium without adenine (Formedium, UK), supplemented with 2% glucose (Sigma-Aldrich, UK) (or 2% galactose (Sigma-Aldrich, UK) for control conditions) inoculated with ~$5x10^8$ cells from population G0. Cells were grown in 500 ml of liquid medium at 30C (or 37C) and 230 RPM for 4 days. To keep the culture between OD600=0.1 and OD600=1.0, ~$5x10^8$ cells were transferred into fresh medium every 12

hours. Five competition experiments were performed, and samples were collected as shown in Table S1. Throughout the manuscript, unless otherwise noted, we show the results of experiment "Small_1_30C_Glu", but the results and conclusions were reproducible between experiments.

During competitive growth, selection could in principle increase the copy number of U3-containing plasmids. This would lead to unreliable fitness estimates, and to overestimation of fitness, particularly for variants whose fitness can be compensated by increased plasmid copy numbers. We estimated the magnitude of these effects by examining variation in fitness estimates among replicate measurements of the same U3 variant, both within and between experiments. Fitness measurements were reproducible, even among low-fitness variants (Figs 1C,D; S3C; S6), suggesting that random effects, such as changes in plasmid copy number, do not play a major role. In addition, copy number changes would lead to overestimation of low fitness effects, and in consequence, a positive bias in epistasis estimates. Thus, selection for copy number cannot explain the observed overall enrichment of negative interactions, nor can it explain any of the other major conclusions of the study.

Sequencing sample preparation

Yeast cells collected at each time point were treated with zymolyase in 3 ml of 20mM KPi pH=7.4 for 1 hour in 37C to remove the cell wall, and plasmid DNA was isolated from remaining spheroplasts using Qiagen MAXIprep, omitting the steps prior to addition of P2 buffer. We amplified 16 ng of template per sample, corresponding to ~ 1.5 x 10^9 plasmid molecules. Illumina (or Ion Torrent) adapters were added in a 25-cycles PCR reaction, performed with Phusion® High-Fidelity PCR Master Mix with HF Buffer (Fisher Scientific, UK) in 50 µl in each of 8 reaction tubes (HiSeq primers: IndexX_PCR_bar_seq with

R_PCR_U3bar_seq; Ion Torrent Proton primers: Proton_trP1_PCR_bar_seq_F with Proton_A_PCR_bar_seq_R and Proton_A_PCR_bar_seq_F with Proton_trP1_PCR_bar_seq_R Table S2). PCR products from all reaction tubes were pooled into 50 µl using a PCR Cleanup column (Qiagen). The appropriate PCR band was isolated by 2% E-Gel SizeSelect agarose gel electrophoresis (Life Technologies, UK) and quantitated by Bioanalyzer (Agilent, IGMM technical support/Edinburgh Clinical Research Facility, UK). For Illumina high-throughput sequencing, we usually pooled 6 samples (for HiSeq) or 2 samples (for MiSeq) mixed in equimolar ratios. In the case of Ion Torrent Proton (IGMM technical support/Edinburgh Clinical Research Facility, UK) we sequenced 1 sample per chip**.**

Counting of barcodes

FASTQ files from 50 bp Illumina HiSeq (or Ion Torrent) sequencing were demultiplexed. The template for deep sequencing of barcodes contained 25 nt fragments flanking the 20 nt barcode sequence of interest. The HiSeq sequencing was performed so that the first position in the read corresponds to the first nucleotide of the barcode and the 3' flanking sequence was found at the 3' end of reads (in the case of Ion Torrent sequencing both flanks were present in the read). We used blast (*30*) and bedtools (*31*) to locate and remove the 3' flanking sequence (or both flanking sequences) from reads and we counted unique barcodes. We recovered barcode sequences that could be uniquely matched to exactly one of the barcode sequences from the filtered list (see "Associating U3 sequences with random barcodes" above), allowing at most two sequencing errors per barcode.

Fitness estimation

In laboratory experiments with continuous growth and overlapping generations, the population size can be represented by the formula $N_t=N_0 \exp(m\ t)$, where N_t is the number of individuals at time t, N_0 is the initial number of individuals, and m is the exponential growth rate, also known as "Malthusian parameter" (*32, 33*). When two or more populations compete with each other, the Malthusian parameter of each population is equivalent to the logarithm of fitness (log fitness), and the difference in Malthusian parameters is equivalent to the difference of log fitness. Throughout this manuscript, we define the relative "log fitness" of a genotype as the Malthusian parameter of that genotype minus the median Malthusian parameter of the wild-type genotypes in the same experiment.

To obtain log fitness estimates from data, we used a Poisson regression approach (*34*) with exponentially decaying mean. Count numbers of barcode *l* at time *t* were modeled as Poisson random variables

$$n_l(t) \sim Po(m_l^t)$$

$$m_l^t = \frac{\exp(\lambda_l t + b_{l0})}{b_t}$$

Here λ_l represents the unknown (un-normalized) log fitness of barcode *l*, and b_t and b_{l0} are normalization factors accounting for different library sizes and different initial counts. We placed Gaussian priors over the λ_l, b_t and b_{l0} variables and obtained Bayesian posterior estimates using the Expectation-Propagation approximation (*35*). Log fitness estimates were then adjusted by subtracting the median log fitness estimate of barcodes corresponding to wild type U3 to produce normalized log fitness estimates.

Reproducibility of fitness measurements for over 2,000 wild-type variants of U3 with different barcodes was used to set the minimal number of reads at G0 which assure reliable results, and we filtered the rest of the barcodes accordingly.

Because mutations known to completely inactivate U3 (in C' and D boxes) all had average log fitness estimates between -2 and -2.5, we reasoned that barcode log fitness estimates below -3 were unreliable. Such estimates showed the largest variation between replicate experiments and were typically based on low numbers of reads. Overall, about 700/22,000 variants in each of the Small library datasets, and 5,000/37,000 variants in the Big library dataset, have log fitness estimates below -3. We therefore replaced these values by -3. Omitting this step increased the noise associated with the data, but did not change any of the conclusions. Replacing this step with a smooth tanh transformation of log fitness estimates, $\lambda'=3*\tanh(\lambda/3)$, with parameters chosen so that the transformation is approximately linear in the [-2,0] range and maps values below -3 to values close to -3, led to the same conclusions.

Positional fitness effects

f_i, the average log fitness effect of mutations in position i in an otherwise wild-type background, was defined as the mean log fitness of variants that had a single substitution or deletion at position *i*, and no other mutations elsewhere.

p_i, the aggregate log fitness effect of position *i* across all genetic backgrounds, was defined as the mean log fitness of variants that had a substitution or deletion at position *i* (and possibly other mutations elsewhere) minus the mean log fitness of variants that had no mutation at position *i*.

Explaining fitness from mutation patterns

In order to attribute the fitness changes to the underlying mutation pattern, we used a regression-based approach. We associate each mutant i with a feature vector z_i, a binary vector whose individual entries correspond to all positions and pairs of positions in the U3 gene. Thus z_i is a vector with 333+(333x332)/2= 55,611 dimensions. z_i will have 1 at entry j<334 if and only if the corresponding U3 variant is mutated in that position. Notice that this is a redundant representation: the binary code in the first 333 entries of z_i uniquely determines the remaining 55,278 entries, nevertheless this redundancy is necessary to disentangle the effects of single mutations from epistatic effects of pairs of mutations.

We then modeled the response variable (log fitness) as a linear function of the corresponding feature vector:

$$\lambda_i = w \cdot z_i + \varepsilon \qquad (1)$$

where w is a weight vector to be learnt from the data and ε is an error term. The weight vector has the same dimensions as the feature vectors z. The weight vector encodes the effect on fitness of the mutations: the first 333 entries code for the effect of single point mutations, while the remaining 55,278 capture the epistatic effects of having a mutation in two different locations.

One can highlight the different type of terms (single point and epistatic terms) by reformulating the response equation (1) in terms of features z_l and z_{lm}, and corresponding weights w_l and w_{lm}. A barcode j would have a non-zero feature l if it was mutated in position l and a nonzero feature lm if it was mutated in both positions l and m. Consequently, the coefficients w_l and w_{lm} would represent the single point and epistatic effects learned by the model. To extract the single point

and epistatic effect terms from the measured fitness values, we solve the following regularized least squares problem

$$w = \arg\min\left(\sum_j \left\|\lambda_j - wz_j\right\|^2 - L(w)\right)$$

where *L(w)* is a regularization term which is needed since the number of parameters which must be estimated exceeds or is comparable to the number of barcodes in each library. We consider two types of regularizers:

$$L(w) = a\sum_j w_j^2$$

so called Tikhonov regularization or ridge regression (RR), avoids overfitting by penalizing large coefficients;

$$L(w) = a\sum_j \left|w_j\right|$$

L1 penalty giving rise to the Lasso regression (*36*), a popular choice (*16*) as it returns sparse estimates where irrelevant coefficients are set to zero. In both cases, the regularization coefficient was chosen by five-fold cross-validation on a grid of values. Out of sample predictions at the optimal regularization value (a=5) indicated a good predictive power, with on average 55% of total test variance explained by the model across the cross-validation runs. These values indicate that the model achieves a good fit without overfitting to the training data. To quantitatively measure the performance of the trained model, we measured the explained variance, i.e. the difference between the initial sample variance and the sum of the squared model residuals, relative to the initial variance. As the distributions of fitnesses and residuals are strongly non-Gaussian, we removed the top and bottom 5% of data to eliminate outliers which

may dominate the estimated variances. Using this procedure, the optimal epistatic model explained 86% of the initial variance, while a nonepistatic model, which used only log fitness effects associated with single mutants, explained 49% of the variance. It should be remarked that using the measured effects of single mutations is not necessarily optimal from the point of view of explained variance. To ensure a fair comparison, we also repeated the fitting procedure using solely features associated with single mutations. This optimal nonepistatic model explained approximately 55% of variance, in line with previous reports (*16*) and considerably below the variance explained by an epistatic model.

We found in our experiments that both Lasso and RR were able to compute reproducible estimates of single-site coefficients, which were in good agreement with the experimentally measured values (Pearson $R=0.87$, $p<0.05$). Estimation of pairwise effects with Lasso was however less successful, as Lasso consistently retained almost exclusively the large epistatic effects associated with pairs of positions with large single-site effects. By contrast, RR captured both positive and negative epistatic effects. Comparisons with directly measured pairwise effects on a subset of mutants with only two mutations reveals that the empirical distribution of pairwise effects appears to encompass both large positive epistatic effects, and smaller (but still important) negative epistatic effects. It therefore appears that Lasso, in the limited data regime we operate, is not well suited to estimate relevant coefficients of highly heterogeneous size, and effectively eliminates most of the negative epistasis by shrinking such coefficients to zero.

Further analysis of the RR results revealed a bias in the estimation of single-site effects (see Fig. S8), which is more pronounced for the Big library (where few single-site mutants were present). This is possibly due to the redundant encoding of mutants, which creates correlated features where explaining away is possible. We therefore further modified the RR model by directly

inputting the w_j coefficients as measured on single-site mutants, and estimated the remaining entries of the w_j vector from the data. The results shown in the manuscript come from this modified model.

The code used in our numerical experiments is available on github: https://github.com/terembura/EpistaticInteractionsYeastU3.

Resolution of fitness estimates

The standard deviation of log fitness estimates of wild-type variants ranges from 0.11 to 0.18 in the small library experiments in glucose. As a result, very small effects, both negative and positive, could not be detected experimentally even if they did play some role in evolution.

There were no single mutants with a fitness estimate greater than 2 standard deviations above wild-type in any experiment. Although 17 single mutants had fitness estimates greater than 1 standard deviation above wild-type, none of these mutants were reproducibly fitter than wild-type in all three replicate experiments. Thus, no mutations increased fitness to a degree detectable with our technology.

Visualization of fitness effects and epistatic interactions

To visualize the fitness effects of individual mutations and their epistatic interactions, we used JavaTreeView (*37*) to generate 2D heatmap plots, VARNA (*38*) to display the effects of mutations along the secondary structure, and Circos (*39*) to generate circular interaction plots. To generate the heatmap shown in Fig. 1D, the fitness effect of each mutation was calculated as the median fitness of single mutant variants that contained the focal mutation, and of the subset of

double mutants that contained the focal mutation plus a low-effect second mutation (|log fitness effect| of the second mutation <0.01). Inclusion of double mutants significantly improved the coverage and reduced noise, without altering the conclusions.

Prediction of U3 secondary structure using epistatic coefficients

To obtain folding constraints for secondary structure prediction, we averaged the epistatic coefficients w_{ij} from four glucose competition experiments, and we filtered these coefficients in two ways. First, we identified all pairs of residues that could potentially interact within uninterrupted stems of 3 or more nucleotides (allowing Watson-Crick and G-U interactions). This retained 7,637 of the 55,278 epistatic interactions. Second, we removed interactions involving at least one large-effect site (defined as regions 80-87 (Box C'), 252-257 (Box C) and 325-329 (Box D). This further reduced the dataset to 6,639 interactions between potentially basepaired sites. We then computed epistatic support scores for each 3-nt stem by averaging the relevant w_{ij} values, as in (*40*). In the resulting set, known interactions had significantly larger support scores than noninteracting pairs (Wilcoxon test, $p<2*10^{-16}$), and in particular, 5 out of 6 top scores corresponded to known basepairing interactions. These 5 scores were used as constraints in RNA structure prediction by the Vienna program (*41*).

Supplementary Figure 1. Mutant library design.

A, B. Uniform coverage of mutations along synthetic genes in the small library (**A**) and big library (**B**).

C. Distributions of the number of single nucleotide substitutions per gene (upper panels) and of substitutions+deletions+insertions per gene (lower panels) in the two mutant libraries.

Supplementary Figure 2. Complementation of D343 strain with U3.

Growth of D343 cells supplemented with wild-type U3 (top panels), empty plasmid (middle panels) or a non-functional mutant of U3 (mutation list: C43T, C75G, A169C, G202T, G248C, C252T, C259T, G282T, A293T, T302A, G302A, G305del; bottom panels) on glucose and galactose media without adenine.

Supplementary Figure 3. Reproducibility of fitness measurements.

A. Reproducibility of fitness measurements for individual barcodes in experiments Small_1_30C_Glu (X axis) and Small_3_30C_Glu (Y axis). R= 0.78.

B. Reproducibility of mean fitness per site (f_i) in experiments Small_1_30C_Glu (X axis) and Small_3_30C_Glu (Y axis). R=0.92.

C. Reproducibility of fitness measurements among high-coverage variants. The mean and standard deviation of log fitness are shown for 824 variants with at least two measurements with 1000 or more reads in population G0 among datasets Small_1_30C_Glu and Small_3_30C_Glu.

Supplementary Figure 4. Effects of mutations on folding energy.

A. Nomenclature of stems in the secondary structure of U3. Yeast-specific regions are shown in blue, regions conserved among Eukaryotes are in red.

B. The mean fitness of U3 variants binned by ascending minimal folding energy, calculated separately for each stem and for hinge–pre-rRNA hybrids. Bin ranges (from left to right, for stems I, II and VI): bin 1, $\Delta\Delta G<0$; bin 2, $\Delta\Delta G=0$; bin 3, $0<\Delta\Delta G<1$; bin 4, $1<\Delta\Delta G<2$; bin 5,

2<ΔΔG<3; bin 6, ΔΔG>3; for stems III, IV, V, 5' hinge and 3' hinge: bin 1, ΔΔG<0; bin 2, ΔΔG=0; bin 3, 0<ΔΔG<2; bin 4, 2<ΔΔG<4; bin 5, 4<ΔΔG<6; bin 6, ΔΔG>6. ΔΔG is the folding energy of the focal mutant minus the folding energy of the wild-type variant.

Supplementary Figure 5. Examples of fitness effects at specific sites.

Fitness effects of each type of mutation (substitutions to A, C, G, T and deletion) at position U191 (A) and positions in terminal stem (B- D).

Supplementary Figure 6. Comparison of fitness effects in 30C and 37C.

A-C, Distribution of p_i, the aggregate log fitness effect of position i across all genetic backgrounds, mapped to the U3 structure for individual experiments: Small_1_30C_Glu **(A)**, Small_3_30C_Glu **(B)**, and Small_2_37C_Glu **(C)**. Arrows indicate positions in which the fitness effects were most different between 30C and 37C.

D-F, Scatter plots of p_i values between pairs of individual experiments.

Supplementary Figure 7. Alternative calculation of the fitness effects of mutations.

The aggregate log fitness effect of each position in experiment Small_1_30C_Glu was calculated as in Fig. S6, except that the median fitness effect of mutations was used instead of the mean, to correct for effects of outliers.

Supplementary Figure 8. Systematic bias of w_i estimated by ridge regression.

A-C, Comparison of single-site effects estimated directly from single mutants (X axis, f_i) and estimated by ridge regression (Y axis, w_i). (A,D), mean of Small_1_30C_Glu and Small_3_30C_Glu; (B,E), Small_2_37C_Glu; (C,F) Big_1_30C_Glu.

D-F, Bias in single-site effects estimated by ridge regression. Regression overestimates f_i for low-effect sites, and underestimates f_i for large-effect sites. To estimate the epistatic effects w_{ij}

by ridge regression, we therefore used empirical estimates of single-site effects from variants mutated at one position only (f_i).

Supplementary Figure 9. Maps of epistatic interactions calculated for individual datasets.

A, mean of Small_1_30C_Glu and Small_3_30C_Glu;

B, Small_2_37C_Glu

C, Big_1_30C_Glu

Supplementary Figure 10. Prediction of U3 secondary structure using constraints from positive epistasis.

A, Map of known basepairing contacts (top left), and of contacts inferred from filtered positive epistatic interactions (bottom right, see Methods). Known contacts are circled.

B, Minimum Folding Energy (MFE) secondary structure predicted by Vienna using U3 sequence data alone.

C, MFE structure predicted by Vienna using U3 sequence and the following basepairing constraints: 330-77 (epistasis score=0.30, epistasis rank=2), 331-76 (score=0.30, rank=3), 332-75 (score=0.23, rank=4), 50-60 (score=0.23, rank=5), 49-61 (score=0.22, rank=6). This structure is markedly more similar to the known structure than structure predicted from sequence alone. In particular, the terminal stem and Stem 2 are now correctly folded. Using centroid instead of MFE structures led to similar results.

Supplementary Figure 11. Cluster of negative epistatic interactions sis in the 3' hinge.

A, Enrichment in negative epistatic interactions in the "3' hinge" (positions 62-73), which mediates basepairing of U3 with pre-rRNA.

B, The mean fitness (left axis) and U3-pre-rRNA interaction energy (right axis) of mutants binned by numbers of mutations in the 3' hinge.

Supplementary Table 1. List of experiments.

Experiment name	Sample ID*	Time (h)†	Number of reads used	Number of barcodes recognised	Number of mutants accepted
Small_1_30C_Glu	total		76505765	41617	22372
	G0	0	14233050		
	D1.25	36	14051955		
	D2.25	60	11881307		
	D3.25	84	15850703		
	D4.25	108	20488750		
Small_2_37C_Glu	total		88384609	40250	23915
	G0	0	14013114		
	D1.25	36	14338456		
	D2.25	60	14277184		
	D3.25	84	14327688		
	D4.25	108	14365043		
	D5.25	132	2699207		
	D6.25	156	14363917		
Small_3_30C_Glu	total		201446365	41266	23163
	G0	0	28745264		
	D0.25	12	28843586		
	D0.75	24	28537727		
	D1.25	36	29034201		
	D2.25	60	28822264		
	D3.25	84	28646717		
	D4.25	108	28816606		
Small_3_30C_Gal	total		165773130	41509	23167
	G0	0	28745264		
	G0.25	12	21662238		
	G1.25	36	28813189		
	G2.25	60	28856414		
	G3.25	84	28866443		
	G4.25	108	28829582		
Big_1_30C_Glu	total		13622683	67069	36692
	G0	0	3488092		
	D0.25	12	2629826		
	D0.75	24	3083151		
	D1.25	36	4421614		

*The Sample ID represents the growth medium (G for galactose, D for glucose), and the duration of selection (in days) that was assumed for that sample in the fitness estimation. For example D1.25 represents a sample that was under selection on glucose for 1.25 days (30 hours). There is a 6-hour difference between the selection time and wall-clock time, because it took approximately 6 hours for genomically encoded U3 to be depleted after addition of glucose.
†Wall-clock time when sample was taken.

Supplementary Table 2. Sequences of oligonucleotides.

Name	Sequence
U3_1F_wt	CTTAAAATCTGTGTCGACGTACTTCATAGGATCATTTCTATAG GAATCGTCACTCTTTGACTCTTCAAAAGAGCCACTGAATCCA ACTTGGTTGATGAGT
U3_1R_wt	ATTGCGGACCAAGCTAATTTAGATTCAATTTCGGTTTCTCACT CTGGGGTACAAAGGTTATGGGACTCATCAACCAAGTTGGA
U3_2F_wt	AAATTAGCTTGGTCCGCAATCCTTAGCGGTTCGGCCATCTATA ATTTTGAATAAAAATTTTGCTTTGCCGTTGCATTTGTAGT
U3_2R_wt	AGTACATAGGATGGGTCAAGATCATCGCGCCATAAAATATTG TAATTACTTCCAAAGGAAAAAACTACAAATGCAACGGCAAA
U3_3F_wt	CTTGACCCATCCTATGTACTTCTTTTTTGAAGGGATAGGGCTC TATGGGTGGGTACAAATGGCAGTCTGACAAGTTAACCAC
U3_3R_wt	TAATCCAATTTCTTAACTGAAAACCAAACCTTTGGTTTTAAAC AATTTAGAAAAGGAAAAAAGTGGTTAACTTGTCAGACT
U3_1F_mut	CTTAAAATCTGTGTCGACG(01019701)(01010197)(97010101)(019 70101)(01010197)(01010197)(01970101)(97010101)(01010197)(97010 101)(01019701)(01019701)(97010101)(01010197)(01970101)(9701010 1)(01010197)(01010197)(01010197)(01970101)(01010197)(97010101) (01010197)(97010101)(01019701)(01019701)(97010101)(97010101)(0 1010197)(01970101)(01019701)(01010197)(01970101)(97010101)(019 70101)(01010197)(01970101)(01010197)(01010197)(01010197)(01019 701)(97010101)(01970101)(01010197)(01970101)(01010197)(0101019 7)(01970101)(97010101)(97010101)(97010101)(97010101)(01019701) (97010101)(01019701)(01970101)(01970101)(97010101)(01970101)(0 1010197)(01019701)(97010101)(97010101)(01010197)(01970101)(019 70101)(97010101)(97010101)(01970101)(01010197)(01010197)(01019 701)(01019701)(01010197)(01010197)(01019701)(97010101)(0101019 7)(01019701)(97010101)(01019701)(01010197)
U3_1R_mut	(97010101)(01010197)(01010197)(01019701)(01970101)(01019701)(0 1019701)(97010101)(01970101)(01970101)(97010101)(97010101)(010 19701)(01970101)(01010197)(97010101)(97010101)(01010197)(01010 197)(01010197)(97010101)(01019701)(97010101)(01010197)(0101019 7)(01970101)(97010101)(97010101)(01010197)(01010197)(01010197) (01970101)(01019701)(01019701)(01010197)(01010197)(01010197)(0 1970101)(01010197)(01970101)(97010101)(01970101)(01010197)(019 70101)(01010197)(01019701)(01019701)(01019701)(01019701)(01010 197)(97010101)(01970101)(97010101)(97010101)(97010101)(0101970

	1)(01019701)(01010197)(01010197)(97010101)(01010197)(01019701)(01019701)(01019701)(97010101)(01970101)(01010197)(01970101)(97010101)(01010197)(01970101)(97010101)(97010101)(01970101)(01970101)(97010101)(97010101)(01970101)(01019701)(01010197)(01010197)(01019701)(01019701)(97010101)
U3_2F_mut	(97010101)(97010101)(97010101)(01010197)(01010197)(97010101)(01019701)(01970101)(01010197)(01010197)(01019701)(01019701)(01010197)(01970101)(01970101)(01019701)(01970101)(97010101)(01010197)(01970101)(01970101)(01010197)(01010197)(97010101)(01019701)(01970101)(01019701)(01019701)(01010197)(01010197)(01970101)(01019701)(01019701)(01970101)(01970101)(97010101)(01010197)(01970101)(01010197)(97010101)(01010197)(97010101)(01010197)(01010197)(01010197)(01010197)(01019701)(97010101)(97010101)(01010197)(97010101)(97010101)(97010101)(97010101)(01010197)(01010197)(01010197)(01010197)(01019701)(01970101)(01010197)(01010197)(01010197)(01019701)(01970101)(01970101)(01019701)(01010197)(01010197)(01019701)(01970101)(97010101)(01010197)(01010197)(01010197)(01019701)(01010197)(97010101)(01019701)(01010197)
U3_2R_mut	(97010101)(01019701)(01010197)(97010101)(01970101)(97010101)(01010197)(97010101)(01019701)(01019701)(97010101)(01010197)(01019701)(01019701)(01019701)(01010197)(01970101)(97010101)(97010101)(01019701)(97010101)(01010197)(01970101)(97010101)(01010197)(01970101)(01019701)(01970101)(01970101)(97010101)(01010197)(97010101)(97010101)(97010101)(97010101)(01010197)(97010101)(01010197)(01010197)(01019701)(01010197)(97010101)(97010101)(01010197)(01010197)(97010101)(01970101)(01010197)(01010197)(01970101)(01970101)(97010101)(97010101)(97010101)(01019701)(01019701)(97010101)(97010101)(97010101)(97010101)(97010101)(97010101)(01970101)(01010197)(97010101)(01970101)(97010101)(97010101)(97010101)(01010197)(01019701)(01970101)(97010101)(01970101)(01019701)(01019701)(01970101)(97010101)(97010101)
U3_3F_mut	(01970101)(01010197)(01010197)(01019701)(97010101)(01970101)(01970101)(01970101)(97010101)(01010197)(01970101)(01970101)(01010197)(97010101)(01010197)(01019701)(01010197)(97010101)(01970101)(01010197)(01010197)(01970101)(01010197)(01010197)(01010197)(01010197)(01019701)(97010101)(97010101)(01019701)(01019701)(01019701)(97010101)(01010197)(97010101)(01019701)(01019701)(01019701)(01970101)(01010197)(01970101)(01010197)(97010101)(01010197)(01019701)(01019701)(01019701)(01010197)(01019701)(01019701)(01019701)(01010197)(97010101)(01970101)(97010101)(97010101)(97010101)(01010197)(01019701)(01019701)(01970101)(97010101)(01019701)(01010197)(01970101)(01010197)(0

	1019701)(97010101)(01970101)(97010101)(97010101)(01019701)(01010197)(01010197)(97010101)(97010101)(01970101)(01970101)(97010101)(01970101)
U3_3R_mut	TAATCCAATTTCTTAACTGAAAACCAAACCTTTGGTTTTAAACAATTTAGAAAAGGAAAAAAGTGGTTA(97010101)(01970101)(01010197)(01010197)(01019701)(01010197)(01970101)(97010101)(01019701)(97010101)(01970101)(01010197)
U3_end_bar_EcoR1	ACGTACGNNNNNNNNNNNNNNNNNNNNTAATCCAATTTCTTAACTGA
U3_start_SalI	TTAAAATCTGTGTCGACG
EcoRI_oligo_SalI	ACGTACGTGAATTCACGTACGTACGTACGTGTCGACACGTACGT
SalI_oligo_EcoRI	ACGTACGTGTCGACACGTACGTACGTACGTGAATTCACGTACGT
Index1_PCR_U3_seq	CAAGCAGAAGACGGCATACGAGATATCATGAGTCAGTCAGCCTGTTTCTACTTAAAATCTGT
Index2_PCR_U3_seq	CAAGCAGAAGACGGCATACGAGATCAAGTTAGTCAGTCAGCCTGTTTCTACTTAAAATCTGT
R_PCR_U3bar_seq	AATGATACGGCGACCACCGAGATCTACACTATGGTAATTGTAAACGACGGCCAGTGAATTC
Index1_PCR_bar_seq	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTCAGTTAAGAAATTGG
Index2_PCR_bar_seq	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTCAGTTAAGAAATTGG
Index3_PCR_bar_seq	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTCAGTTAAGAAATTGG
Index4_PCR_bar_seq	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTCAGTTAAGAAATTGG
Index5_PCR_bar_seq	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTCAGTTAAGAAATTGG
Index6_PCR_bar_seq	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTCAGTTAAGAAATTGG
Custom_Read1_seq_p	TATGGTAATTGTAAACGACGGCCAGTGAATTC

rimer_U3	
Proton_A_PCR_bar_seq_F	CCATCTCATCCCTGCGTGTCTCCGACTCAGGTTTTCAGTTAAGAAATTGG
Proton_A_PCR_bar_seq_R	CCATCTCATCCCTGCGTGTCTCCGACTCAGTTGTAAAACGACGGCCAGTG
Proton_trP1_PCR_bar_seq_F	CCTCTCTATGGGCAGTCGGTGATGTTTTCAGTTAAGAAATTGG
Proton_trP1_PCR_bar_seq_R	CCTCTCTATGGGCAGTCGGTGATTTGTAAAACGACGGCCAGTG

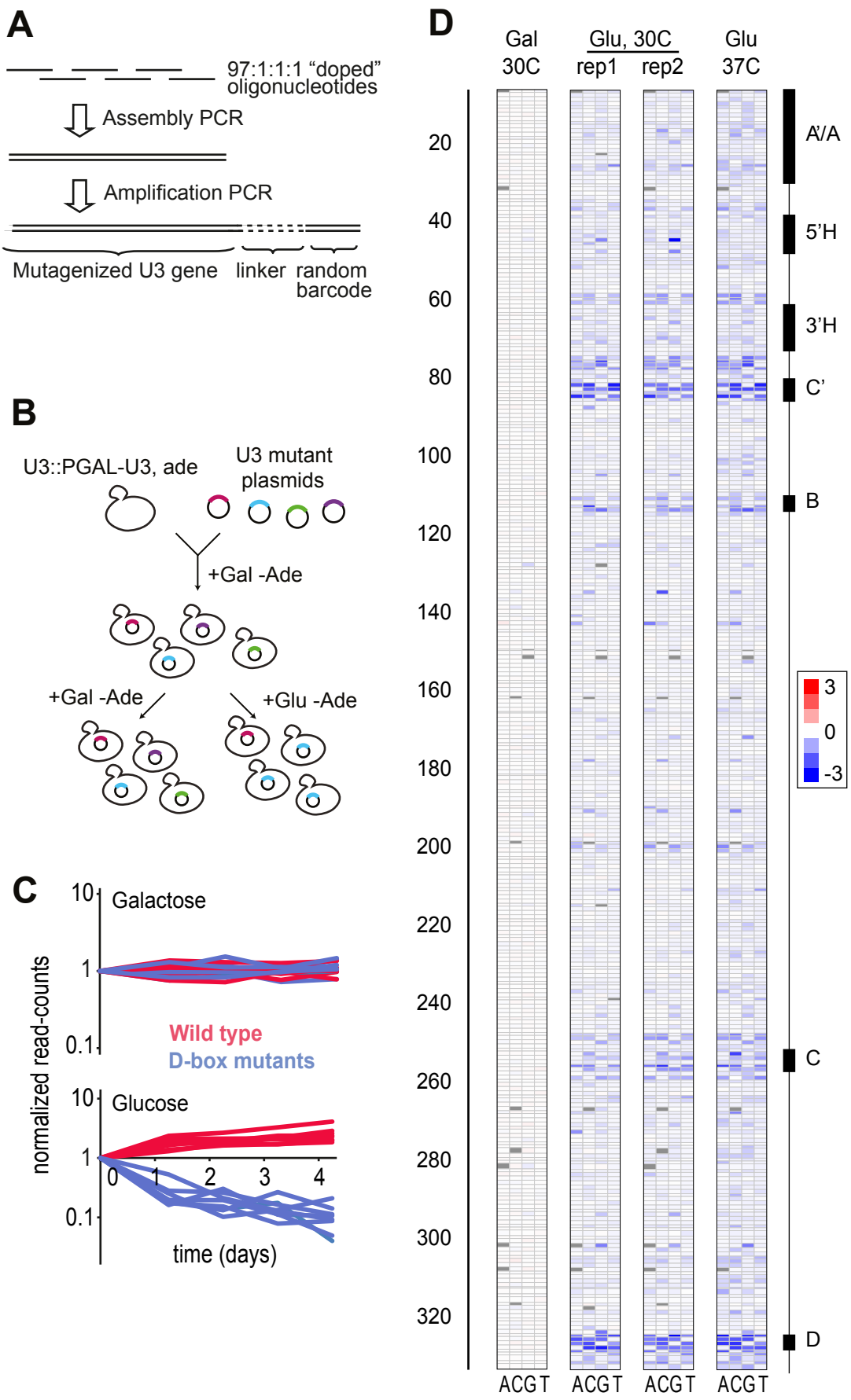


Figure 1
Puchta et al.

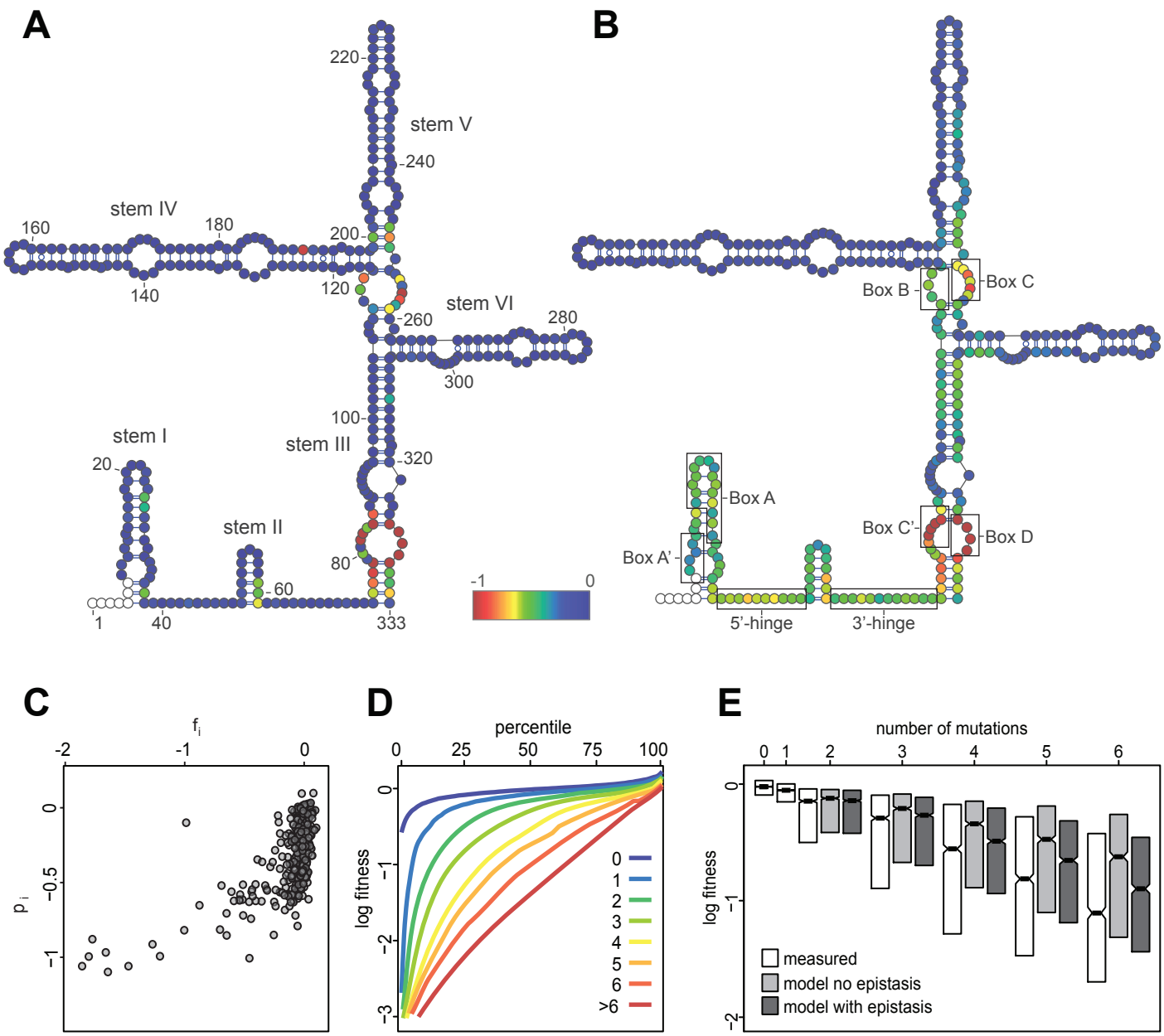


Figure 2
Puchta et al.

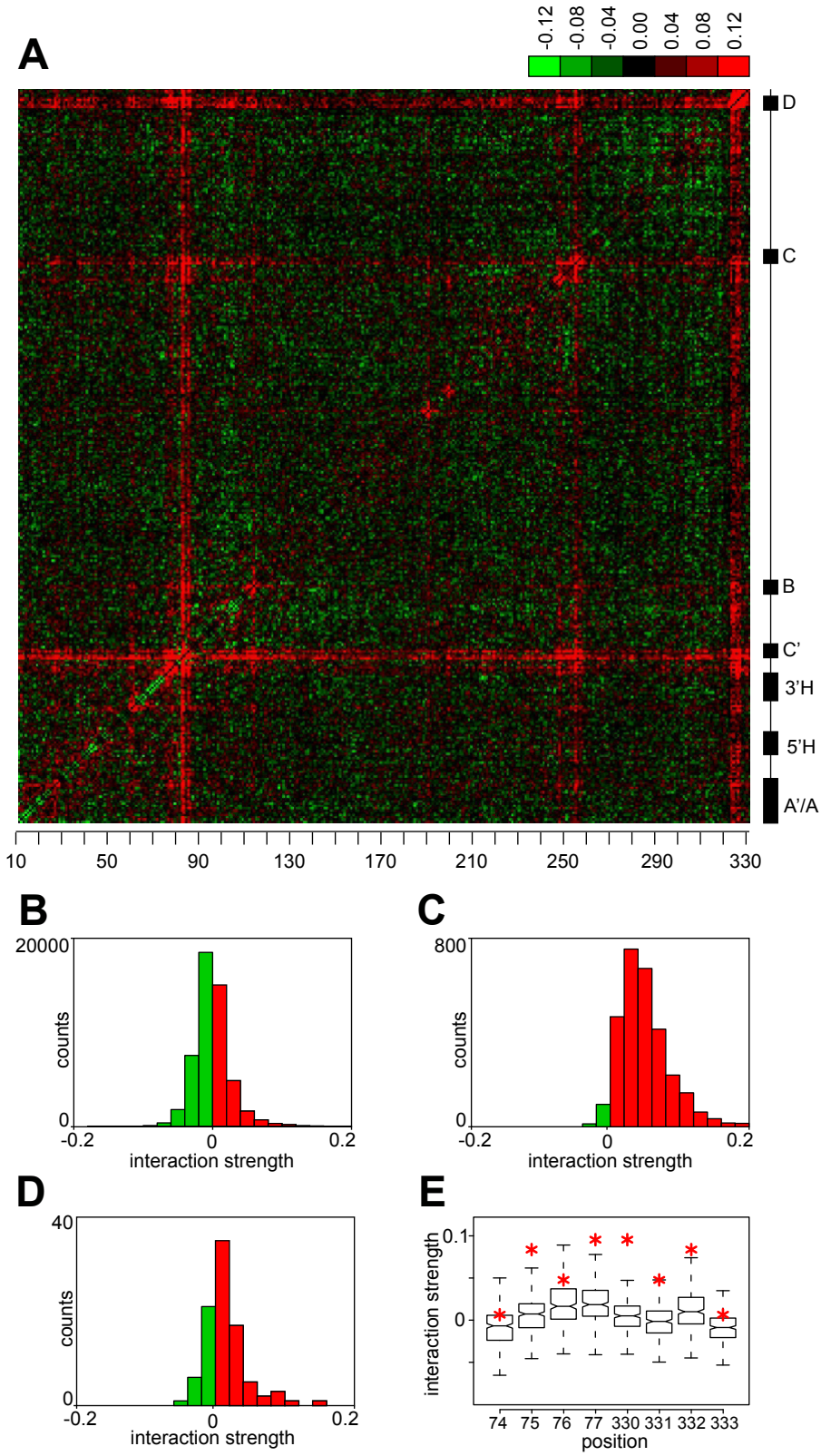


Figure 3
Puchta et al.

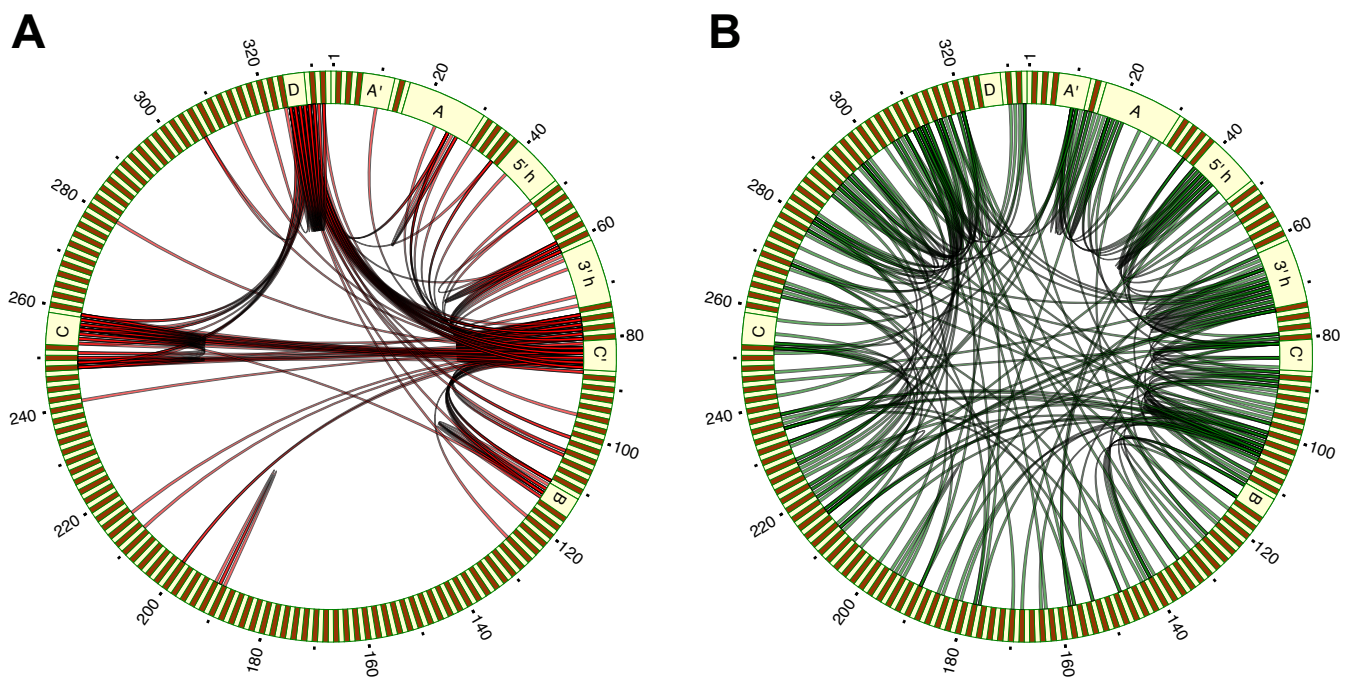
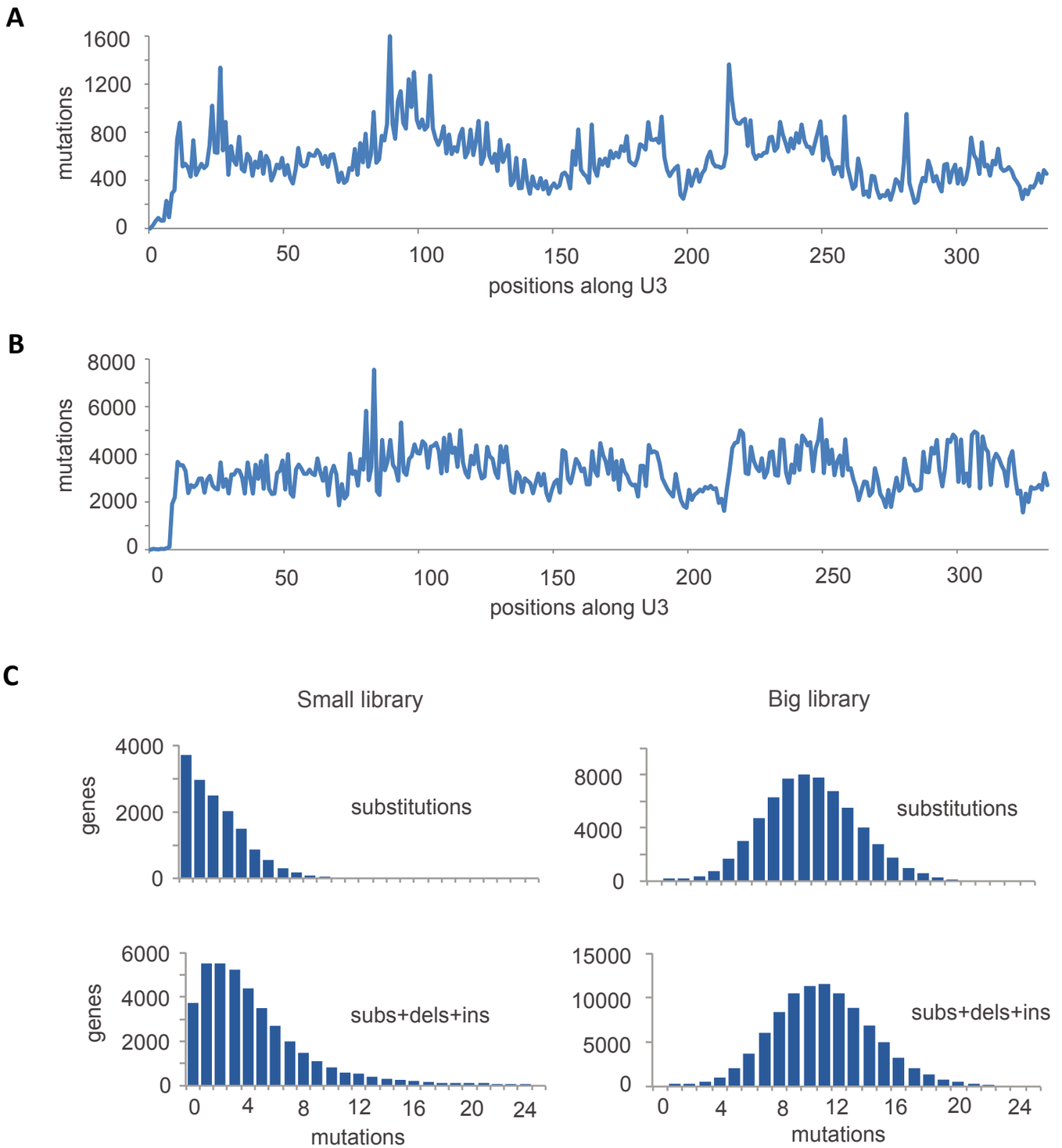
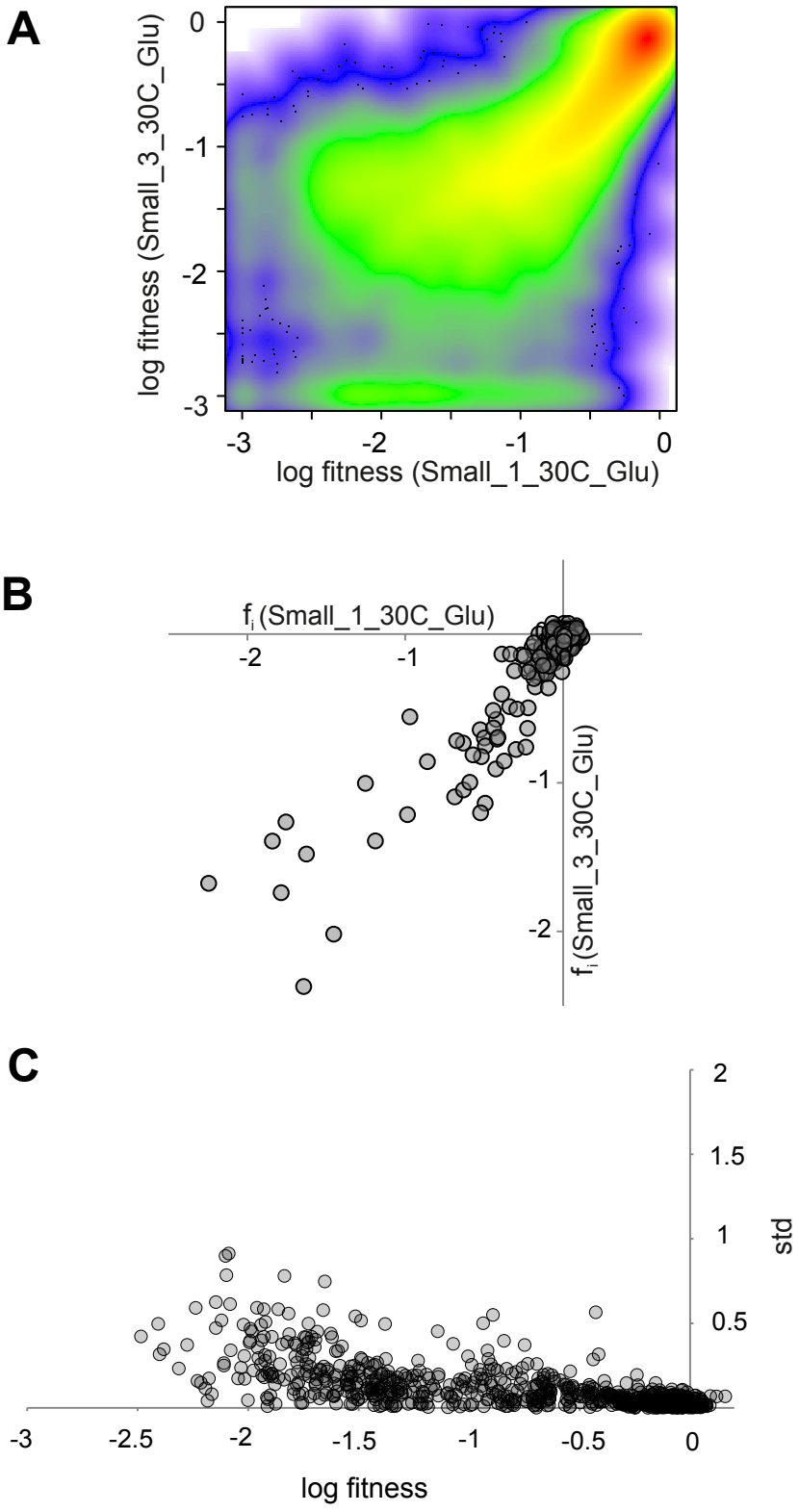


Figure 4
Puchta et al.

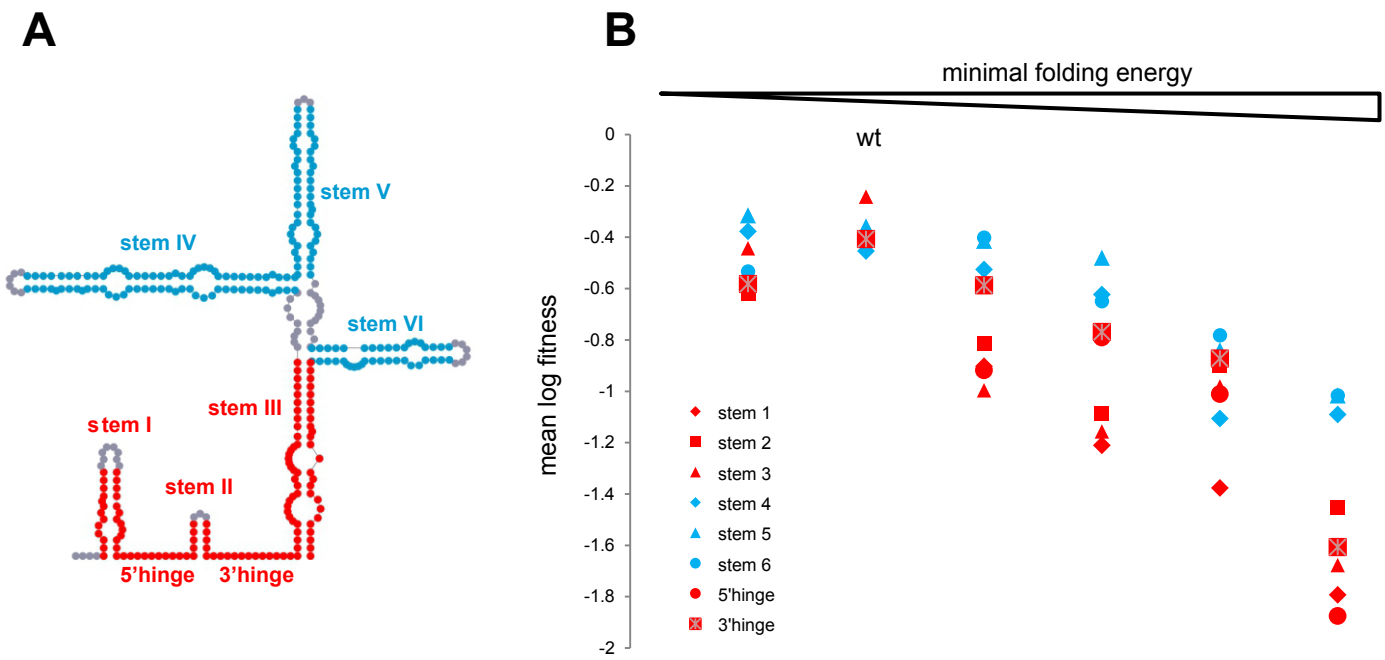


Supplementary Figure 1
Puchta et al.

Supplementary Figure 2
Puchta et al.

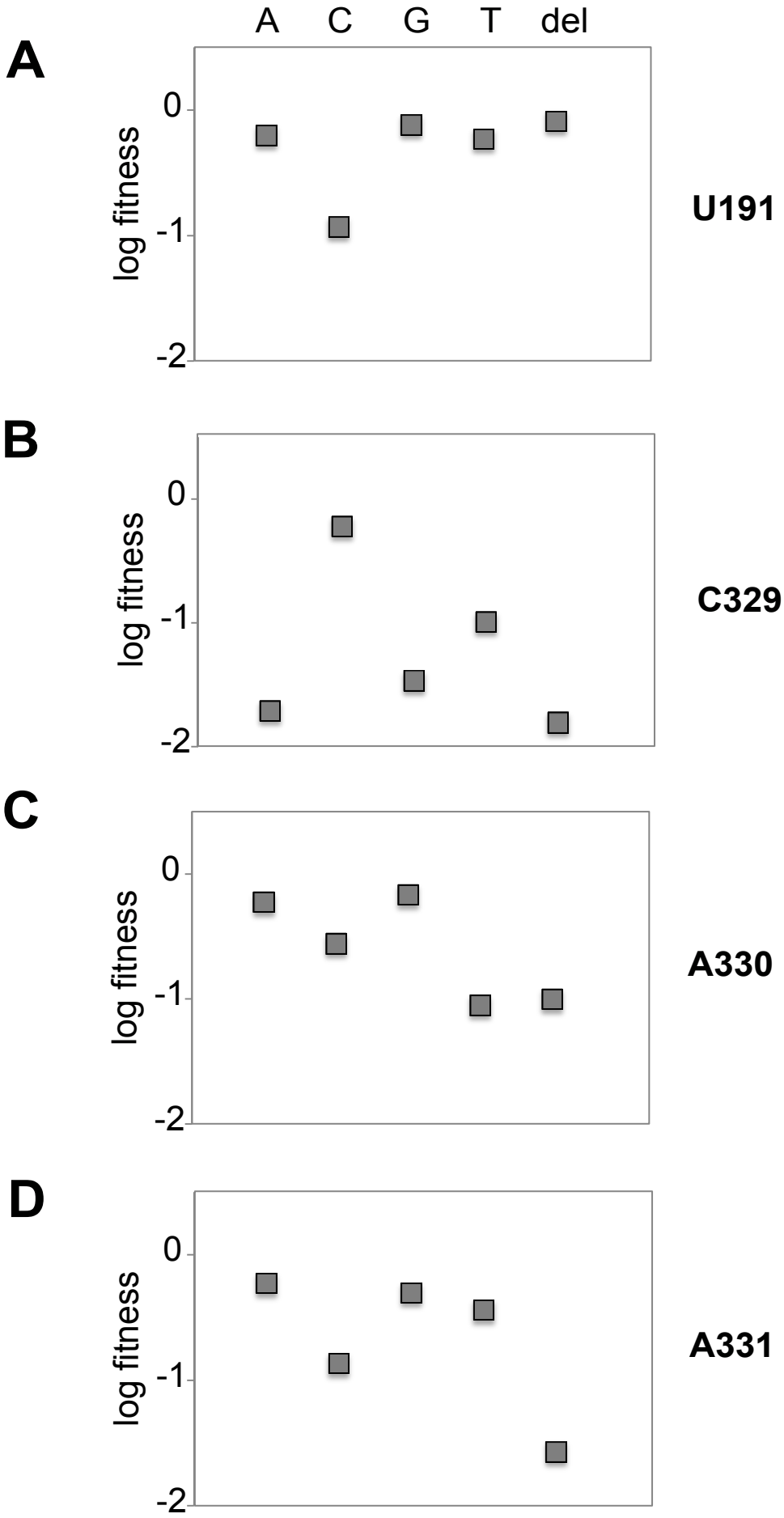


Supplementary Figure 3
Puchta et al.

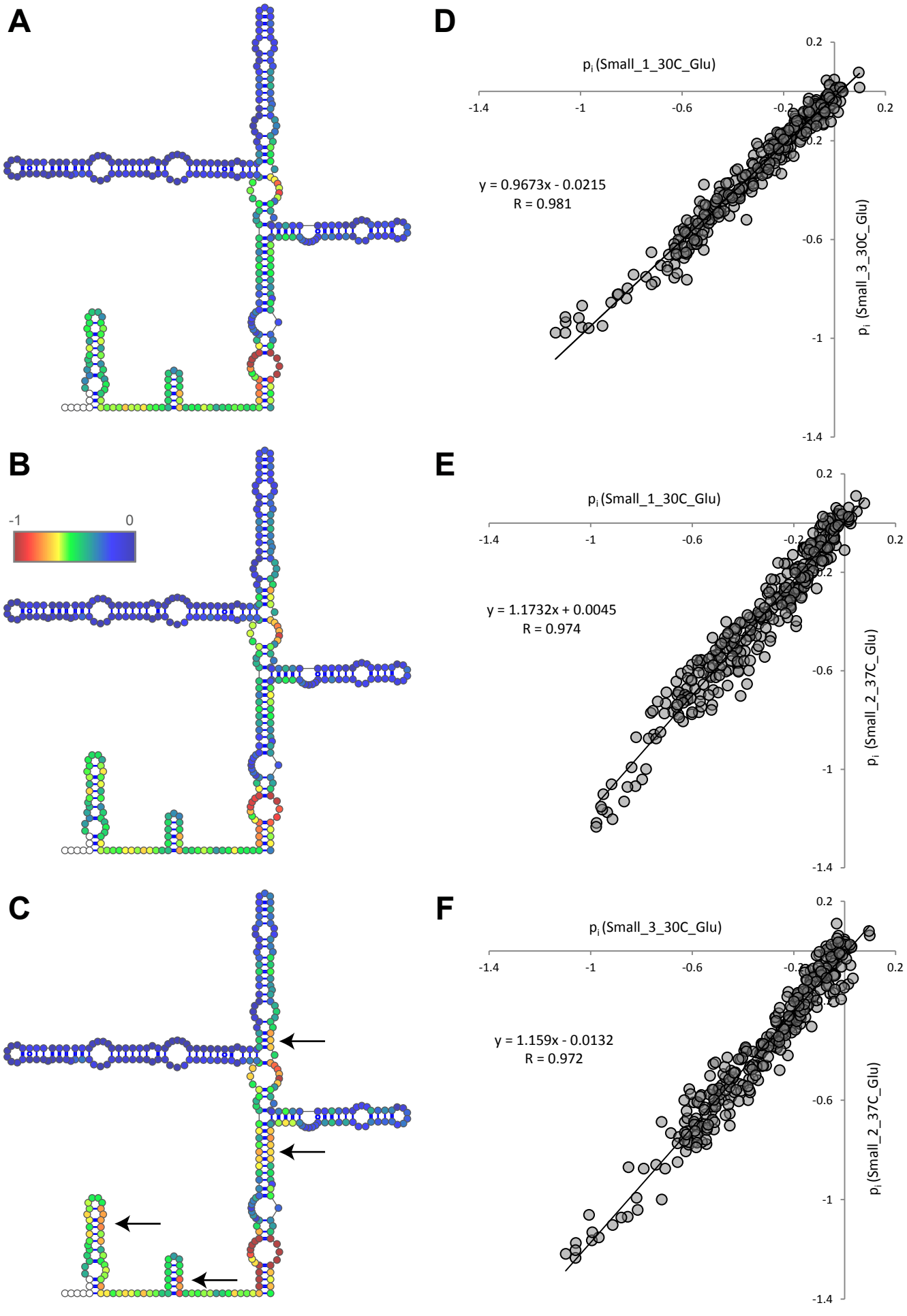


Supplementary Figure 4

Puchta et al.

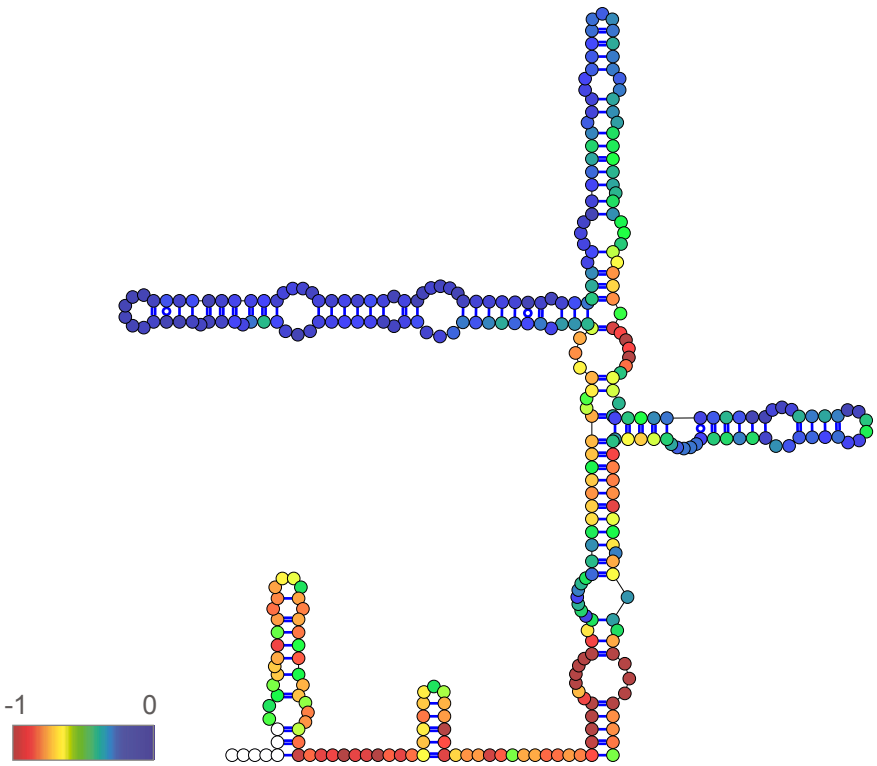


Supplementary Figure 5
Puchta et al.



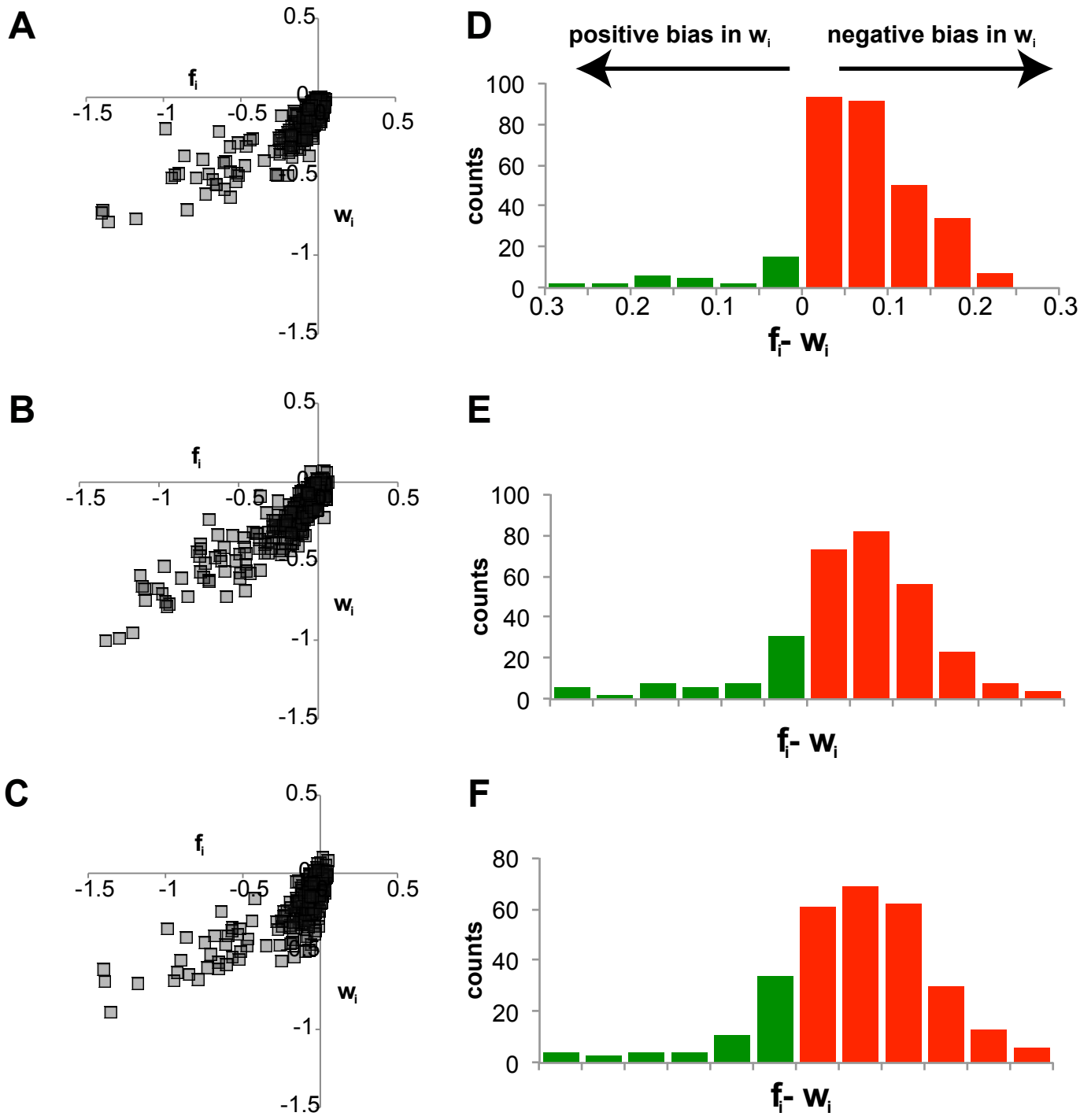
Supplementary Figure 6

Puchta et al.

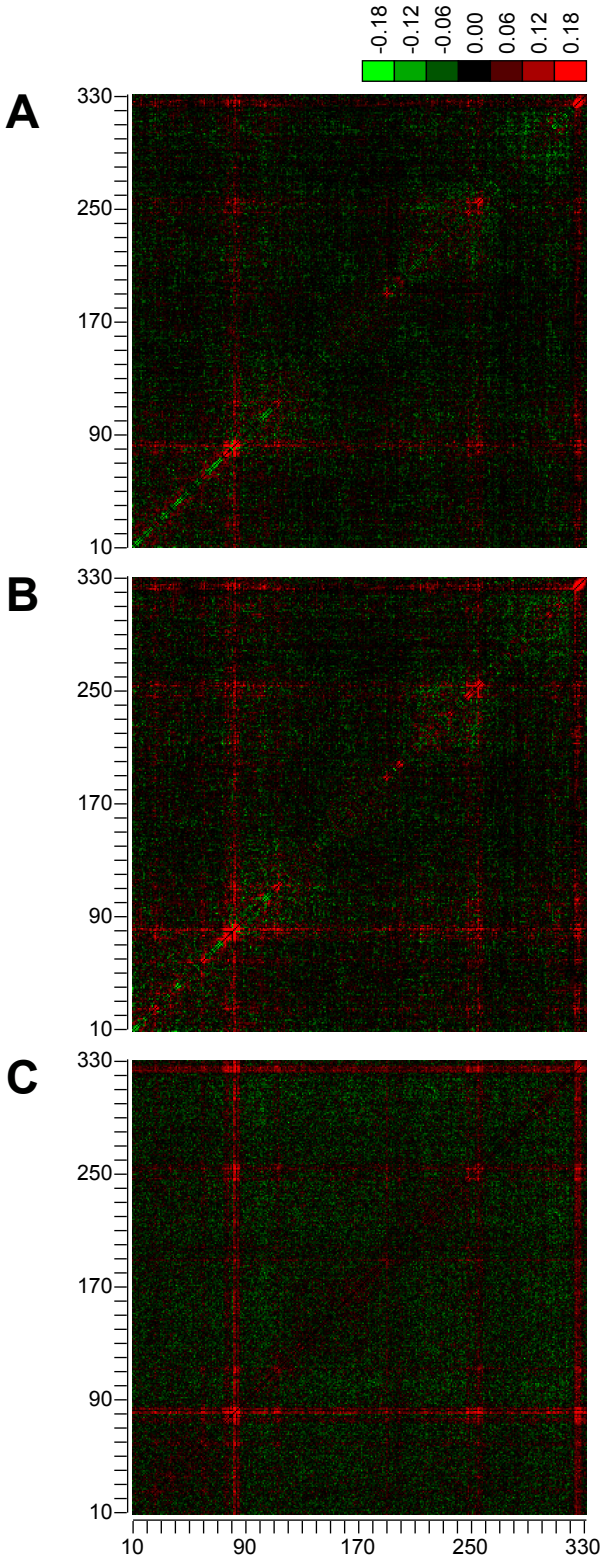


Supplementary Figure 7

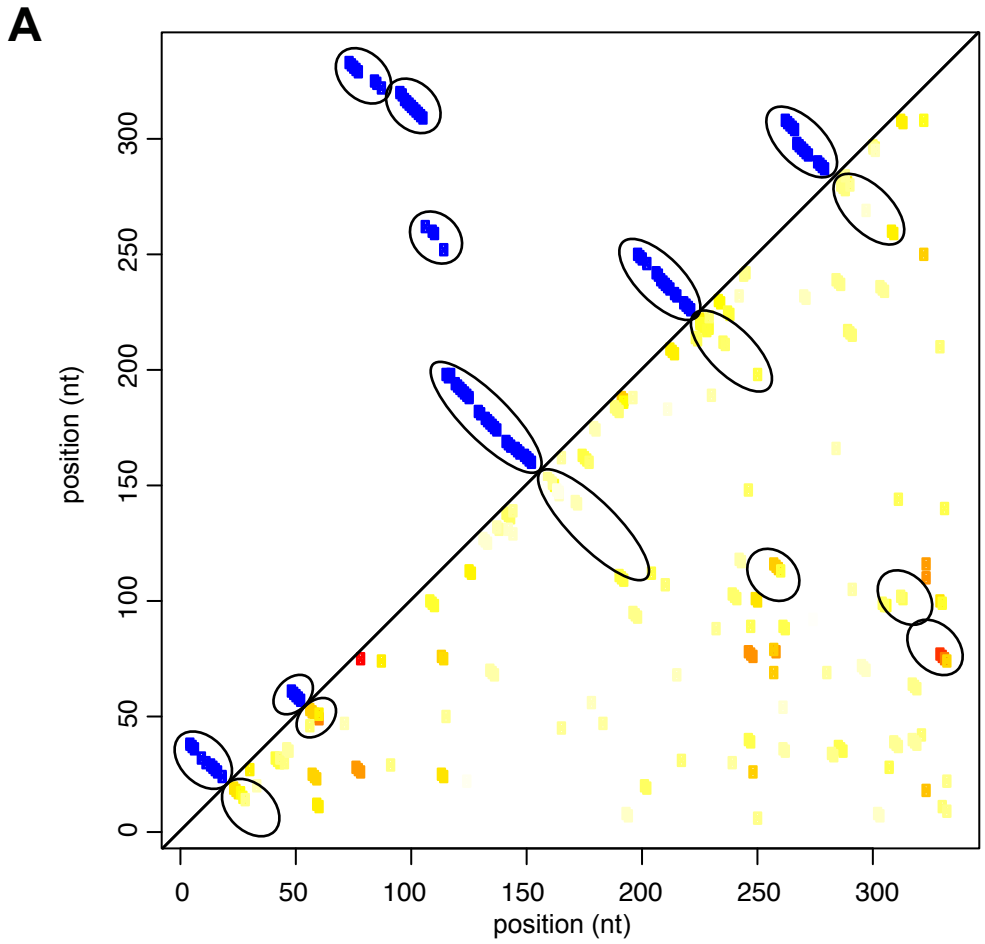
Puchta et al.

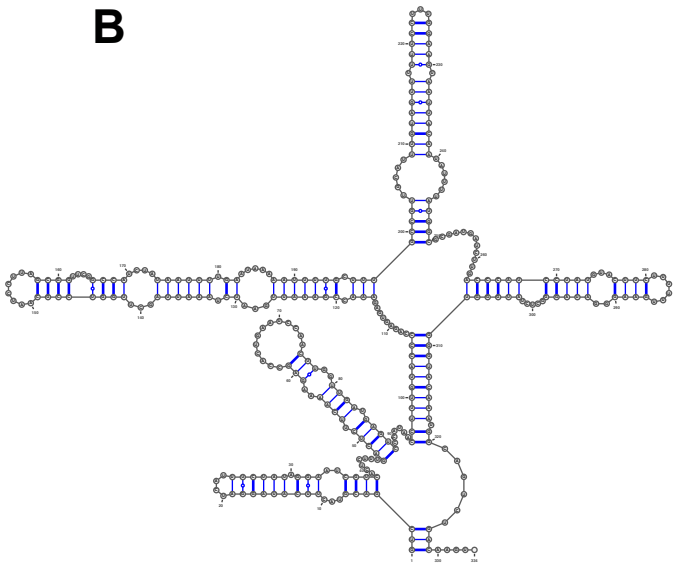


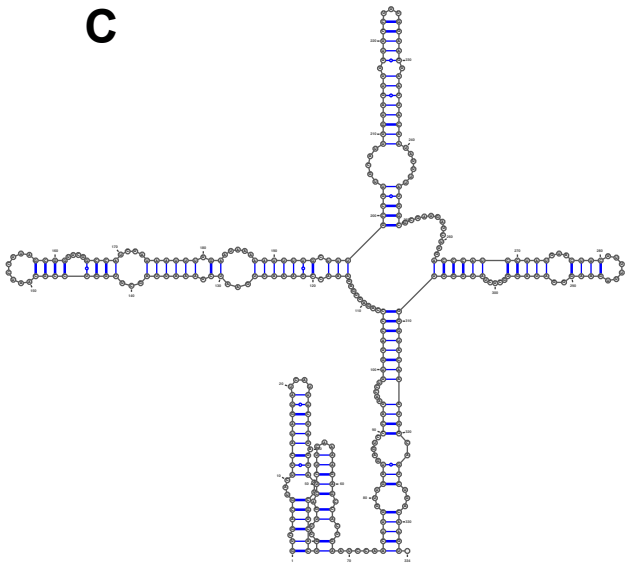
Supplementary Figure 8
Puchta et al.



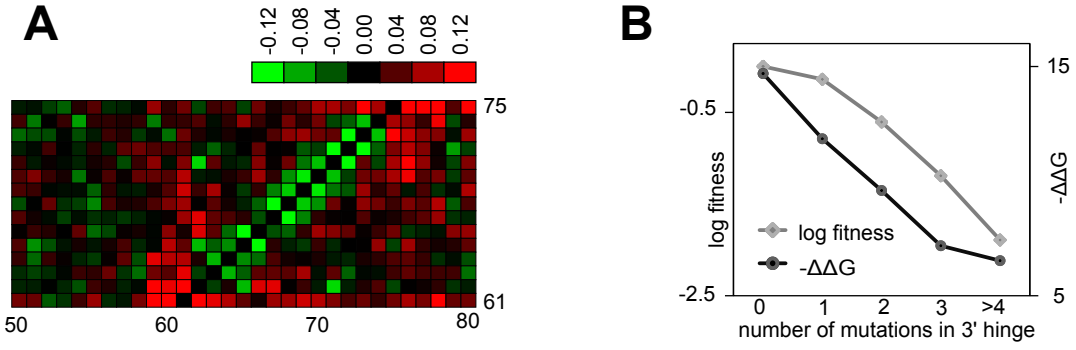
Supplementary Figure 9
Puchta et al.







Supplementary Figure 10
Puchta et al.



Supplementary Figure 11
Puchta et al.